STROKE KNOWLEDGE GRAPHS · RELATION EXTRACTION · LLM-AS-A-JUDGE

Constructing and Evaluating a Stroke Knowledge Graph

From extraction-method comparison to the fidelity of LLM-as-a-judge

Elena Atanasovska and Boshko Koloski

boshko.koloski@ijs.si
robin.lemmens@uzleuven.be
carlosav.molina@vallhebron.cat
pietro.caliandro@policlinicogemelli.it
marko.robniksikonja@fri.uni-lj.si
dragi.kocev@ijs.si

Combining two studies: *KG Construction Methods on the Stroke Domain* (NFMCP 2025)

The stroke burden — and why a knowledge graph

A “silent pandemic” whose literature has outgrown human synthesis

2nd
leading cause of death worldwide

9.7M
projected annual deaths by 2050 (+50%)

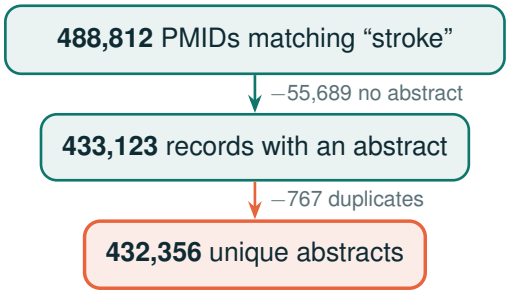
$2.3T
projected global economic toll by 2050

86%
of stroke deaths fall on low- & middle-income countries

The literature grows by **tens of thousands of stroke abstracts a year** — too vast for any team to read and connect. **Knowledge graphs** turn this unstructured text into queryable **(subject, predicate, object)** facts, enabling multi-hop queries, link prediction, and clinical decision support.

The Stroke-PubMed corpus

A large-scale, domain-specific foundation built from PubMed



- Queried via the NCBI **E-utilities** API; abstracts through **April 2025**.
- Spans **8,940 journals** and **over 1M** unique authors.
- Experiments run on a **representative 10k subset**, word-count matched to the full corpus.

Part 1

Constructing the StrokeKG

Comparing four relation-extraction paradigms

Four extraction paradigms

Spanning unsupervised, supervised, and generative approaches

OpenIE

RULE-BASED · UNSUPERVISED

Open schema, no training. High triplet volume, but low precision.

REBEL

SUPERVISED · SEQ-TO-SEQ

End-to-end generation; full coverage, fixed 169-relation schema.

ReLiK

SUPERVISED · RETRIEVE-AND-LINK

High-precision linking, bound to a fixed 193-relation schema.

Gemma 2 9B

LLM · FEW-SHOT

Role-plays a stroke expert; open schema, filters for relevance.

The LLM-as-a-judge evaluation framework

Ten clinical criteria — later reviewed & endorsed by three stroke-medicine specialists

GPT-4o-mini scores each triplet **1 (poor) – 5 (excellent)** on every criterion.

Criterion	w	Criterion	w
Clinical Relevance	2.1	Diagnostic Utility	1.7
Evidence Strength	2.3	Therapeutic Implications	2.1
Specificity	1.4	Prognostic Value	2.3
Guideline Concordance	2.1	Population Impact	2.1
Pathophys. Accuracy	2.1	Potential for Harm	1.9

COMPOSITE QUALITY SCORE

$$\mathrm{CQS} = \sum_{i=1}^{10} w_i \, s_i$$

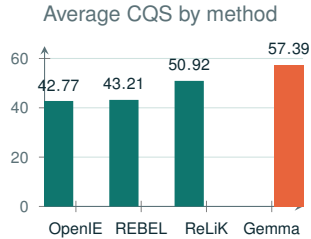
Range **20.1 – 100.5**

A scalable proxy — not a replacement for expert validation.

Construction results: volume vs. quality

The most prolific method is the least clinically useful

Method	Triplets	Uniq. rel.	Select.	Avg CQS
OpenIE	327,404	52,174	0.0%	42.77
REBEL	142,760	169	0.0%	43.21
ReLiK	51,426	193	6.6%	50.92
Gemma 2 9B	**72,591**	**13,095**	**13.9%**	**57.39**



Selectivity = share actively filtered out as irrelevant at extraction time.

OpenIE: maximal volume, minimal quality · **REBEL/ReLiK:** capped by tiny fixed schemas · **Gemma:** highest quality *and* intelligent on-the-fly filtering.

What each method extracts

From granular schema facts to high-level clinical synthesis

Subject	Predicate	Object	Method
Mechanical thrombectomy	has use	large vessel occlusion	ReLiK
ischemic stroke	subclass of	stroke	REBEL
Reperfusion therapy	is treatment for	acute ischemic stroke	OpenIE
Oral anticoagulation	reduces	risk of stroke	Gemma
Atrial fibrillation	causes	cardioembolic stroke	Gemma
hypertension	increases	stroke risk	Gemma

TAKEAWAY. For specialized domains, the **open-schema, contextual** extraction of LLMs wins — Gemma tops the most heavily weighted criteria (**Clinical Relevance**, **Pathophysiological Accuracy**, **Potential for Harm**), while supervised models excel at precise schema facts and OpenIE captures both facts and noise.

Part 2

Can we trust the judge?

The construction study assumed the LLM judge was reliable — is it?

Two questions the construction study left open

It ranked extractors with one LLM judge — and proposed surrogate scorers as future work

RQ1 | RELIABILITY

How **stable** are LLM quality judgments across **models**, **batch sizes**, and **temperatures**?

If scores shift with configuration, the assessment itself becomes an uncontrolled variable.

RQ2 | SCALABILITY

Can **small, fine-tuned models** match a **proprietary LLM judge** — scaling to the full corpus?

Exactly the surrogate-scorer idea the construction study proposed for the 432k corpus.

Two complementary contributions: a consistency analysis of LLM judges + surrogate distillation into a cheap, deterministic scorer.

Experimental design

One prompt, 24 scoring configurations, the same triplets throughout

- **100 triplets** sampled from the held-out test split
  (28 OpenIE, 31 REBEL, 22 ReLiK, 19 Gemma).
- **5 OpenAI models:** GPT-4, GPT-4o-mini, GPT-4.1-mini, GPT-5-mini, GPT-5.
- **3 batch sizes** (1, 5, 20 triplets per prompt).
- **Temperature** $\in \{0.2, 1.0\}$ where exposed.
- Prompt template held constant — only *model*, *batch*, *temperature* vary.

5 models $\times$ 3 batch sizes $\times$ temperature

24

distinct scoring configurations

$\Rightarrow$ up to 24 scores per triplet–criterion pair

Finding 1: LLM judges are not stable

Across all 1,000 triplet–criterion groups

2.47

average per-triplet score range (on the 1–5 scale)

2.0%

of groups reach *full* agreement; 14.3% near ($\leq$1)

44%

average pairwise *exact* agreement between configs

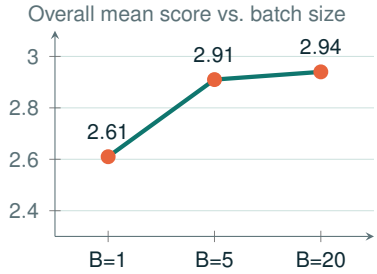
2.34–3.09

model mean scores: GPT-5 strictest, GPT-4 most generous

Even the same model at different batch sizes disagrees — e.g. GPT-4.1-mini agrees with itself on only **63%** of scores between batch 1 and batch 20.

The prompt-context effect dominates

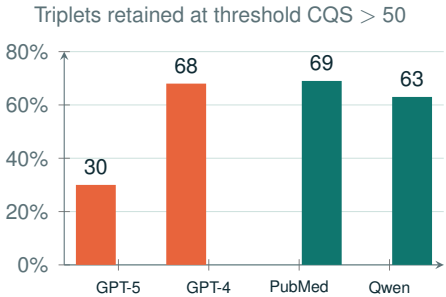
Presenting more triplets per prompt systematically inflates scores



- In **52.3%** of triplet–criterion groups, batch-size variance *exceeds* inter-model variance.
- Friedman test: significant prompt-context effect for **9 of 10** criteria (Potential for Harm the exception).
- Distributional shift, not just a mean shift: modal score moves from **1** (B=1) to **3** (B=20).
- Likely driven by **anchoring**, **implicit comparison**, and **context-length** effects — a prompt-context effect, not a pure batch artifact.

Why instability matters downstream

The judge you pick changes the size — and composition — of the resulting KG



GPT-5 / GPT-4 are LLM judges; PubMed / Qwen are surrogates.

- Per-triplet CQS range averages **35.9 points** on the 20.1–100.5 scale.
- At CQS > 50, retention swings from **30% (GPT-5)** to **68% (GPT-4)** — the same corpus yields a KG **more than twice as large**.
- Conservative newer models penalize valid-but-*underspecified* facts — changing what the KG *contains*, not just its size.

Cross-study comparison needs strict configuration control.

Part 3

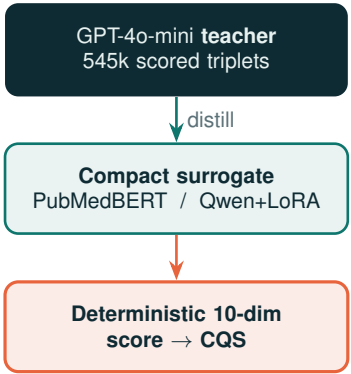
Scaling the evaluation

Distilling one LLM judge into a compact, deterministic surrogate

Surrogate distillation

Lock in one teacher's scoring as a deterministic, auditable artifact

- **Teacher:** GPT-4o-mini scores on **545,645** triplets (80/10/10 split).
- Train two surrogates as **multi-task regressors** on the 10-dim score vector (MSE loss):
  - **PubMedBERT** — 110M params, full fine-tune
  - **Qwen2.5-1.5B** — LoRA, only **18.5M** trainable (1.18%)
- Single NVIDIA A100; deterministic, continuous-valued predictions.



Surrogates faithfully replicate the judge

High teacher fidelity at a tiny fraction of the cost

Metric	Baseline	PubMed	Qwen
CQS Pearson r	—	0.815	**0.825**
CQS Spearman ρ	—	0.794	**0.802**
CQS MAE ↓	17.61	9.28	**8.50**
MSE ↓	1.365	0.616	**0.575**

Held-out test set (54,565 triplets). Both surrogates *halve* the baseline error.

Deployment on 545k triplets (single A100)

Scoring time	5 days → **15–50 min**
Inference cost	$35 → **$1–$3**
Speedup	**144–480×**
Determinism	**guaranteed**

Qwen edges out PubMedBERT while updating just **1.18%** of its parameters.

An honest caveat

Fidelity to the teacher is not the same as clinical validity

- Correlations measure **teacher fidelity** — how well surrogates *copy* GPT-4o-mini — not agreement with clinical ground truth.
- Surrogates **inherit the teacher's biases**: they make one (possibly flawed) evaluation cheaper, faster, deterministic — they do not make it *correct*.
- The architecture is **teacher-agnostic**: retrain on expert labels, no changes needed.

HARDEST CRITERION

Potential for Harm

$r \approx$ **0.48–0.51**

Also the highest cross-config variance among judges — partly a *noisy teacher signal*, capping achievable fidelity.

Part 4

The road ahead

From a calibrated judge to a continuously-built, validated StrokeKG

Next steps

Refine the graph, validate it predictively, and keep it live

a **Calibration triplets.** Extract a representative sample and have the three stroke experts score it — gold labels to calibrate the judge and re-train the surrogates.

b **Ontology distillation.** Distill a consolidated ontology of entity and relation types, taming open-schema sprawl (e.g. OpenIE's 52k relations).

c **Ontology-guided rewrite.** A second LLM pass canonicalizes each triplet to the ontology — normalizing entities and predicates into a linkable graph.

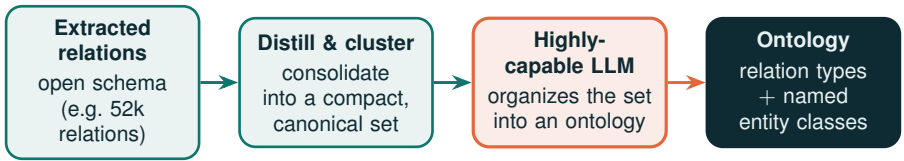
d **Temporal completion evaluation.** Build the KG from literature up to year Y, then predict facts for Y+1 . . . now via KG completion — validating against what actually emerged.

e **Online KG building.** Move from batch rebuilds to incremental, streaming construction that ingests new abstracts continuously.

Steps a–c refine the graph · d validates it · e keeps it live.

Ontology distillation

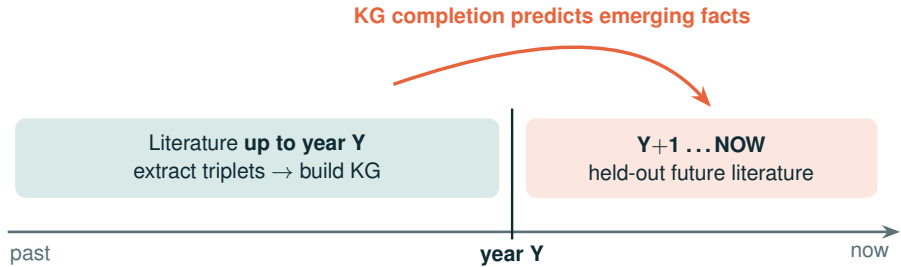
From open-schema relations to a clean, LLM-generated ontology



HOW IT WORKS. We feed the **extracted relations** into a distillation step that clusters and consolidates them into a compact, canonical set. A **highly-capable LLM** then organizes this set into an **ontology** — defining the relation types and the **named entity classes** they connect — giving the graph a consistent schema that the ontology-guided rewrite (step c) and KG completion (step d) can rely on.

Validating the graph itself: temporal completion

Can the KG predict the discoveries that came after its cutoff?



WHY IT MATTERS. A forward-looking, leakage-free *validity* test that complements teacher fidelity: if the KG's predicted links match facts that genuinely emerged after year Y, the graph captures real, generative structure — not merely how the LLM judge would have scored it.

The combined picture

- **Construction:** an LLM extractor (Gemma 2 9B) gives the best clinical quality — open schema and intelligent filtering beat rule-based volume and supervised fixed schemas.
- **Reliability:** but the LLM *judge* is unstable across models, batch sizes, and temperatures — any deployment must **freeze a configuration**.
- **Scalability:** compact surrogates replicate the judge deterministically and **144–480**$\times$ cheaper (CQS Pearson ≈ 0.82).
- **Validity:** fidelity $\neq$ clinical validity — the payoff depends on calibration against expert judgment.

The vision: an expert-calibrated, ontology-grounded StrokeKG — validated by temporal completion and built continuously, online.